ELSEVIER

International Biodeterioration & Biodegradation 52 (2003) 243–246

INTERNATIONAL BIODETERIORATION & BIODEGRADATION

www.elsevier.com/locate/ibiod

On analytical methods and inferences for 2 × 2 contingency table data using wildlife examples

Richard M. Engeman[a,*], George D. Swanson[b]

[a] USDA/Wildlife Services, National Wildlife Research Center, 4101 LaPorte Ave, Fort Collins, CO 80521-2154, USA
[b] Physical Education and Exercise Science Department, California State University, Chico, CA 95929-0300, USA

Abstract

The 2 × 2 contingency table is a common analytical method for wildlife studies, but inappropriate analyses and inferences are not uncommon. Issues of concern are presented for selecting the appropriate test for analyzing these data sets. These include the choice of test relative to experimental or sampling design and breadth of intended inferences, the careful statement of hypotheses, and analyses with small sample sizes. Examples from the wildlife literature are used to reinforce the statistical concepts.
Published by Elsevier Ltd.

Keywords: Chi-square (χ^2); Fisher's exact test; McNemar's test; Small sample

1. Introduction

Nominal data that can be expressed as 2 × 2 contingency tables of counts are common in wildlife studies. These data sets are usually subjected to seemingly elementary statistical analyses and most investigators probably feel competent to analyze the data. Reinforcing this, most statistical program packages offer easy access to a variety of tests for analyzing these data sets. Many statisticians consider the analyses of 2 × 2 tables to be trivial and may pass that notion along to their clients. However, the appropriate analysis for a 2 × 2 contingency table often is not trivial. Special attention should be paid to small sample size data sets that are most likely to result in an inappropriate analysis.

The best known and most commonly applied statistic for analyzing 2 × 2 contingency tables is probably the Pearson χ^2. This test is exactly equivalent to the z-test for comparing two binomial proportions, making its use even more commonplace. The Yates (1934) correction for continuity is often applied, even though it long has been recognized to produce a very conservative test resulting in unduly large p-values, especially for small sample sizes (Grizzle, 1967; Upton, 1982; D'Agostino et al., 1988). The continuity corrected statistic commonly appears with the Pearson χ^2 statistic in computer programs. Fisher's "exact" test (e.g., Irwin, 1935) also appears in program outputs for analyzing 2 × 2 tables, although it may not be available for the more general $r \times c$ tables.

Expressing data in a 2 × 2 contingency table does not express the experimental or sampling design. Without knowing the experimental design, 2 × 2 data from independent samples appear identical to those from studies with matched or paired observations. The investigator must be aware of the differences in data structure and hypotheses, and that a test such as McNemar's test (e.g., Sokal and Rohlf, 1981) should be applied to account for the paired data structure.

The experimental or sampling design and inferences of interest must dictate the statistical test to apply. However, rules of thumb as to which test to apply abound for small sample size situations, but advice from different sources is not always consistent. Applied statistics texts frequently instruct that Fisher's "exact" test should be applied when certain criteria are not met for small sample sizes. For example, Dixon and Massey (1969) recommend using the χ^2 statistic only if all expected cell frequencies are greater than or equal to 2, whereas Snedecor and Cochran (1980) say to use Fisher's exact test if the total sample size is less than 20 or if the total sample size is between 20 and 40 and the smallest expected cell frequency is less than 5. The application of this test has become increasingly controversial (e.g., Berkson, 1978). Upton (1982) indicates that Pearson's χ^2 performs better at smaller sample sizes than indicated in

* Corresponding author. Tel.: +1-970-266-6091; fax: +1-970-266-6089.
E-mail address: richard.m.engeman@aphis.usda.gov (R.M. Engeman).

0964-8305/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/S0964-8305(03)00111-2

most texts, whereas Fisher's test has been shown to be very conservative (e.g., Upton, 1982; D'Agostino et al., 1988). Many textbooks advise that Fisher's test should still be used when sample sizes are too small for the Pearson χ^2. Similarly, when the criteria for application of Pearson's χ^2 are not met, many computer program outputs will flag those results and recommend the use of Fisher's test.

In the next three sections, we illustrate our concerns about applying the appropriate test for 2 × 2 contingency table data, how the hypothesis of interest influences the test to use, and some considerations for the small sample situation. We use information from three avian studies to illustrate our points. These examples indicate the care needed for analyzing these "simple" contingency tables and we hope they serve to increase awareness of the potential for problems when analyzing these data.

2. Choice of test

Characteristics of the experimental design and intended inferences determine the most appropriate test to apply. The data in Table 1 originates from a study designed to assess whether tape recorded alarm calls enhanced detectability of Cooper's hawks (*Accipiter cooperii*) (Rosenfield et al., 1988). Observations were made on each sampling transect both with and without recorded calls. Thus, the observations are paired rather than independent. Nevertheless, Rosenfield et al. (1988) must have assumed that the observations under the two conditions from each transect could be considered independent. Their analysis for the nestling stage data (Table 1) using Fisher's test resulted in a one-tailed p-value of 0.08. They concluded, using the data in Table 1 and some additional data, that during the nestling stage broadcast recordings "can markedly increase the chance of detecting Cooper's hawks near their nests." However, an appropriate analysis would have been McNemar's test for paired data.

Besides selecting an appropriate test the analysis of these data can be used to illustrate other issues. As a second point, let us presume that for some reason we could assume that observations with and without recorded calls were independent rather than paired. Then the data set, by all criteria of which we are aware, still is of sufficient size to apply Pearson's χ^2, rather than applying Fisher's test as the authors have done. Lastly, a one-tailed p-value of 0.08 by itself would not be considered by many to indicate a strong (marked) increase in the chance of detecting the Cooper's hawks. If we ignore that McNemar's is the more appropriate test and for illustrative purposes apply Pearson's χ^2, then the one-tailed p-value reached by this analysis would have been 0.04, which could more easily be described as supporting a statement that broadcast calls "markedly" increase the chance of detection. However, we note the more appropriate analysis using McNemar's test results in a one-tailed p-value of 0.08. Thus, the authors coincidentally arrived at the correct p-value through an incorrect application of Fisher's test.

3. Importance of hypotheses

A careful statement of the hypothesis to be tested is important for determining the correct test to apply to the data. As an example, we consider data on the use of nest boxes by Barrow's goldeneyes (*Bucephala islandia*) Savard (1988). In Table 2 we attempt to reproduce 1 of the 2 × 2 data sets, although we were not able to exactly duplicate his χ^2 test statistic reported for these data. Each nest in Table 2 received repeated observations: whether it was successful in year 1 and again in the following year. We assume that each nest was observed for only two consecutive years. The data are paired data, as are some of the other data sets described by Savard. The appropriate analysis for the data in Table 2 is dictated by the hypothesis of interest. Savard desired to know if nesting success in the second year was independent of success in the first year. The population of nests can be considered as divided into the two independent subpopulations of successful and unsuccessful first year nests. An appropriate analysis for testing the independence of the second year nest success rates from these distinct subpopulations is Pearson's χ^2, rather than McNemar's test for paired data. Savard used a χ^2 test to arrive at a p-value of 0.016. We presume that the test used was Pearson's χ^2, because we obtained a similar two-tailed p-value of 0.013 when we applied that test.

As another illustration, a seemingly slight change in the hypothesis of interest dictates a different analysis. Suppose the hypothesis of interest was to compare first

Table 1
Data for effectiveness of broadcast calls for detecting Cooper's hawks at the nestling stage, taken from Rosenfield et al. (1988)

Use of tapes	Detected		Total no. of transects
	Yes	No	
Yes	9	9	18
No	4	14	18
Total	13	23	36

Table 2
Second year nesting success for successful and unsuccessful nests in the first year, Savard (1988)

Success in first year	Success in second year		
	Yes	No	Total
Yes	46	35	81
No	34	56	90
Total	80	91	171

year success rates to second year success rates. Then the population of nests is not divided into two distinct subpopulations, but rather is a collection of paired observations on each nest. Assuming this was the hypothesis of interest, we applied McNemar's test (e.g., Sokal and Rohlf, 1981), which produced a 1 degree of freedom χ^2 statistic with a resulting p-value of 0.904, a result quite different than that given by Pearson's χ^2. Although we had difficulty in duplicating Savard's results exactly (because we could not determine the data used), his explicit definition of the hypothesis to be tested allowed us to determine that an appropriate analysis suited to his hypothesis was performed.

4. Alternatives for small samples

A variety of approaches have been suggested for analysis of 2×2 tables with small sample sizes. Rands and Hayward (1987) compared survival and chick production of wild gray partridges (Perdix perdix) versus hand-reared and released gray partridges (Table 3). First, they compared over-winter disappearance rates among sexes and arrived at a p-value of < 0.05. By duplicating their analysis and comparing our test statistic to theirs, we concluded that they performed an Yate's continuity corrected χ^2 with a two-tailed p-value of 0.003. Second, they compared breeding success of hand-reared and wild pairs of partridges. Here too, they arrive at a p-value < 0.05. We concluded that Pearson's χ^2 was applied with a resulting two-tailed p-value of 0.015.

The authors did not explain that they used the continuity corrected χ^2 in the first analysis and the usual (Pearson) χ^2 in the second. When the SAS PROC FREQ (SAS/STAT User's Guide, 1990) is used to analyze these data sets, a warning about small cell size appears for analyzing both data sets. The zero cell in the first data set results in a small expected frequency in the cell and, therefore, poses a validity problem for applying both Pearson's χ^2 and the continuity corrected χ^2. Both data sets represent situations where investigators are frequently led to Fisher's test, by following the sample-size warnings from software packages. For these data sets, the conservative nature of Fisher's test does not result in an inferential problem because the two-tailed p-values are 0.002 and 0.023 for the first and second data sets in Table 3, respectively.

Table 3
Survival and breeding success data for hand-reared gray partridges Rands and Hayward (1987)

Survival:	Disappearance overwinter		
Sex	Yes	No	Total
Male	10	4	14
Female	0	9	9
Total	10	13	23
Breeding:	Success		
Raised	Yes	No	Total
Hand	7	7	14
Wild	16	2	18
Total	23	9	32

Pearson's χ^2 (or the z-test for comparing two proportions) probably work well at surprisingly small sample sizes, although there is no consensus of opinion as to the exact minimal sample or cell size requirements for a valid test. If sample size is a concern, one could consider an alternative analysis. We illustrate by applying the unconditional test by McDonald and Milliken (1975), McDonald et al. (1977). This test does not seem to be well known, although it is frequently referenced in articles on analyzing 2×2 tables. However, this test is not incorporated into standard program packages and the user must rely on published tables to conduct the test. We applied McDonald's test to these data and the two-tailed results were $p < 0.017$ for the first data set and $p < 0.047$ for the second (for these sample sizes see the tables in McDonald and Milliken (1975)). We emphasize that McDonald's test is not the only test developed to be valid and sensitive for small sample 2×2 contingency tables, but it is easy to use because its results are tabulated up to sample sizes where the Pearson χ^2 is valid. Good reviews of possible tests to apply are given in D'Agostino et al. (1988) and Upton (1982), with the later paper giving a comparative overview of many analytical methods.

5. Discussion

The investigator must assume responsibility for assuring that the correct analysis and inferences are produced from a study. A quick perusal of almost any wildlife journal issue will indicate the strong reliance on data from 2×2 tables in wildlife research. Inappropriate analyses are common, which compromise the biological inferences. The articles from which we obtained our examples were selected because they provided the data structures and analytical applications to illustrate our points, although we noted several places where analyses and inferences could have been strengthened. The analysis of 2×2 data is greatly facilitated by the ready availability of computer software to handle these data. However, this benefit can lead to problems if the investigator is not familiar with other analytical methods which may not be contained in the available software package. Prior to conducting an analysis, the investigator must understand the interplay of (1) the hypothesis of interest, (2) the experimental design, and (3) the limitations that the structure of the data can impose on the use of an analytical method.

Acknowledgements

We wish to thank W. Dusenberry, M. O'Connell, D. Otis and an anonymous referee for their very helpful reviews of this paper.

References

Berkson, J., 1978. In dispraise of the exact test. Journal of Statistical Planning and Inference 2, 27–42.

D'Agostino, R.B., Chase, E., Belanger, A., 1988. The appropriateness of some common procedures for testing the equality of two independent binomial proportions. American Statistician 42, 198–202.

Dixon, W.J., Massey Jr., F.J., 1969. Introduction to Statistical Analysis, 3rd Edition. McGraw-Hill, New York.

Grizzle, J.E., 1967. Continuity correction in the χ^2 test for 2×2 tables. American Statistician 21, 28–32.

Irwin, J.O., 1935. Tests of significance for differences between percentages based on small numbers. Metron 12, 83–94.

McDonald, L.L., Milliken, G.A., 1975. A nonrandomized test for comparing two proportions. College of Commerce and Industry Research Paper 94, University of Wyoming, Laramie, Wyoming.

McDonald, L.L., Davis, B.M., Milliken, G.A., 1977. A nonrandomized unconditional test for comparing two proportions in 2×2 contingency tables. Technometrics 19, 145–157.

Rands, M.R., Hayward, T.P., 1987. Survival and chick production of hand-reared gray partridges in the wild. Wildlife Society Bulletin 15, 456–457.

Rosenfield, R.N., Bielefeldt, J., Anderson, R.K., 1988. Effectiveness of broadcast calls for detecting Cooper's hawks. Wildlife Society Bulletin 16, 210–212.

SAS Institute, Inc., 1990. SAS/STAT User's Guide. SAS Institute, Cary North Carolina.

Savard, J.L., 1988. Use of nest boxes by Barrow's goldeneyes: nesting success and effect on the breeding population. Wildlife Society Bulletin 16, 125–132.

Snedecor, G.W., Cochran, W.G., 1980. Statistical Methods, 7th Edition. Iowa State University Press, Ames, IA.

Sokal, R.R., Rohlf, F.J., 1981. Biometry. W.H. Freeman & Co., San Francisco.

Upton, G.J.G., 1982. A comparison of alternative tests for the 2×2 comparative trial. Journal of the Royal Statistical Society, Series A 145, 86–105.

Yates, F., 1934. Contingency tables involving small numbers and the χ^2 test. Journal of the Royal Statistical Society (Suppl.) 1, 217–235.